Классификация возраста морских улиток (Abalone) на основе физических измерений

Выполнила Жиденко Виктория Александровна
Группа М8О-307Б-23

Задача бинарной классификации

Определить возраст морской улитки (молодая / старая)

по 7 физическим измерениям и полу

Постановка задачи

• Вход: длина, диаметр, высота, вес (общий, без панциря, внутренностей, панциря), пол

• Выход: бинарная метка target = 1, если Rings > 10 (старая улитка)

• Основная метрика: F1-score (учитываем дисбаланс классов)

• Цель: построить модель с высокой обобщающей способностью и реализовать интерактивный калькулятор

Базовая статистика по датасету до обработки

	count	mean	std	min	25%	50%	75%	max
Length	4177.0	0.523992	0.120093	0.0750	0.4500	0.5450	0.615	0.8150
Diameter	4177.0	0.407881	0.099240	0.0550	0.3500	0.4250	0.480	0.6500
Height	4177.0	0.139516	0.041827	0.0000	0.1150	0.1400	0.165	1.1300
Whole_weight	4177.0	0.828742	0.490389	0.0020	0.4415	0.7995	1.153	2.8255
Shucked_weight	4177.0	0.359367	0.221963	0.0010	0.1860	0.3360	0.502	1.4880
Viscera_weight	4177.0	0.180594	0.109614	0.0005	0.0935	0.1710	0.253	0.7600
Shell_weight	4177.0	0.238831	0.139203	0.0015	0.1300	0.2340	0.329	1.0050
Rings	4177.0	9.933684	3.224169	1.0000	8.0000	9.0000	11.000	29.0000

Предобработка и фичи

Предобработка

• One-Hot кодирование: Sex → Sex_F, Sex_I, Sex_M

• Бинаризация: Rings > 10 → target (1 — старая)

• Удаление Rings

Генерация признаков (13 новых)

• volume = Length × Diameter × Height

• density = Whole_weight / volume

• shell_ratio = Shell_weight / Whole_weight

• shell_per_gram = Shell_weight / (Whole_weight − Shell_weight)

• log_whole_weight, density_sq, shell_ratio_sq, sqrt_whole_weight и др.

Ключевой признак: Shell_weight — сильная корреляция с возрастом

Базовая статистика по датасету после обработки

	count	mean	std	min	25%	50%	75%	max
Length	4175.0	0.524065	0.120069	0.075000	0.450000	0.545000	0.615000	0.815000
Diameter	4175.0	0.407940	0.099220	0.055000	0.350000	0.425000	0.480000	0.650000
Height	4175.0	0.139583	0.041725	0.010000	0.115000	0.140000	0.165000	1.130000
Whole_weight	4175.0	0.829005	0.490349	0.002000	0.442250	0.800000	1.153500	2.825500
Shucked_weight	4175.0	0.359476	0.221954	0.001000	0.186250	0.336000	0.502000	1.488000
Viscera_weight	4175.0	0.180653	0.109605	0.000500	0.093500	0.171000	0.253000	0.760000
Shell_weight	4175.0	0.238834	0.139212	0.001500	0.130000	0.234000	0.328750	1.005000
Sex_F	4175.0	0.313054	0.463792	0.000000	0.000000	0.000000	1.000000	1.000000
Sex_I	4175.0	0.320958	0.466901	0.000000	0.000000	0.000000	1.000000	1.000000
Sex_M	4175.0	0.365988	0.481764	0.000000	0.000000	0.000000	1.000000	1.000000
length_diameter_ratio	4175.0	1.291867	0.059066	0.493333	1.257732	1.288462	1.321839	2.333333
log_whole_weight	4175.0	0.568058	0.268286	0.001998	0.366204	0.587787	0.767094	1.341689
volume	4175.0	0.034732	0.021145	0.000041	0.017944	0.032886	0.048777	0.205137

Базовая статистика по датасету после обработки

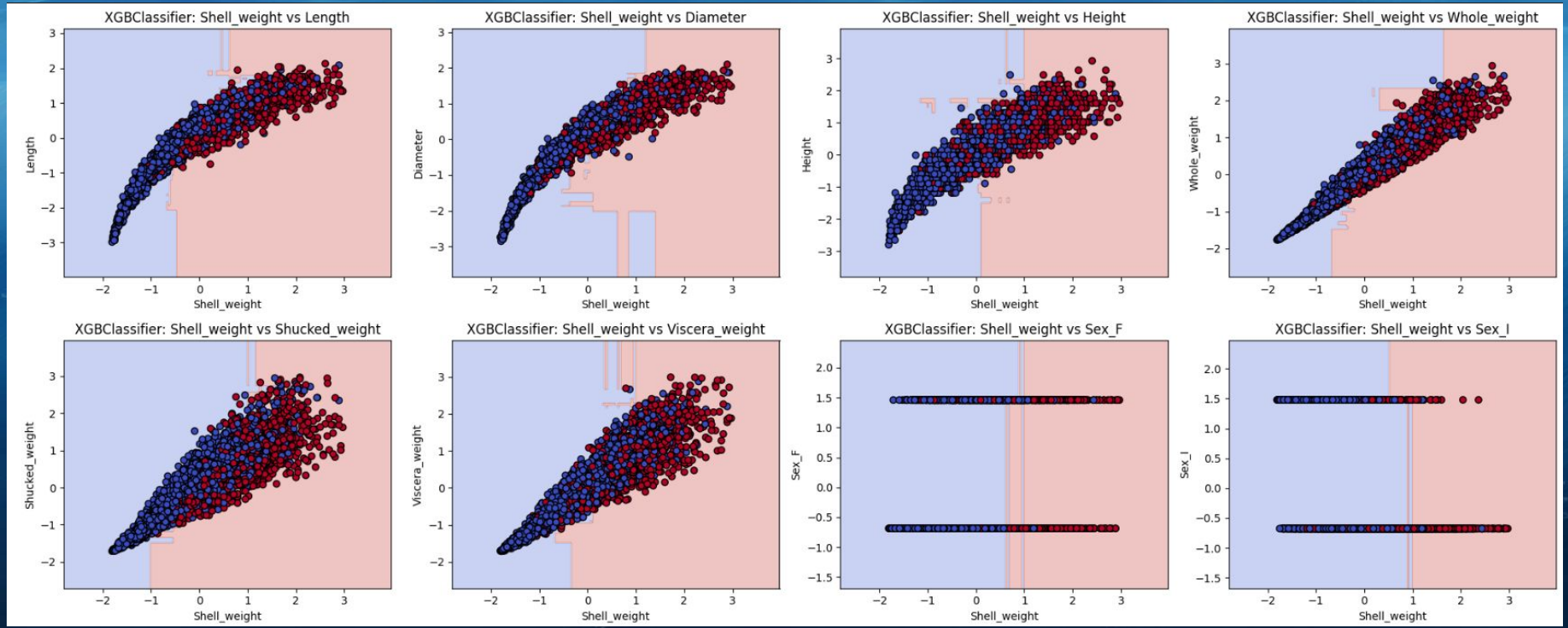
density	4175.0	24.589943	5.220748	3.254380	22.224486	24.082668	26.226240	245.287521
shucked_ratio	4175.0	0.432403	0.105787	0.175258	0.395093	0.430592	0.466170	4.691943
shell_ratio	4175.0	0.295055	0.046555	0.109341	0.266073	0.290870	0.319410	0.935361
shell_per_gram	4175.0	0.429512	0.262218	0.122764	0.362533	0.410178	0.469313	14.468886
growth_efficiency	4175.0	0.041702	0.007352	0.004077	0.038130	0.041524	0.044995	0.307278
viscera_ratio	4175.0	0.218542	0.034368	0.007634	0.198584	0.217259	0.236949	0.665399
density_sq	4175.0	631.914955	951.349283	10.590989	493.927793	579.974910	687.815647	60165.967899
shell_ratio_sq	4175.0	0.089225	0.033645	0.011955	0.070795	0.084605	0.102023	0.874901
growth_efficiency_sq	4175.0	0.001793	0.001543	0.000017	0.001454	0.001724	0.002025	0.094420
sqrt_whole_weight	4175.0	0.863366	0.289177	0.044721	0.665019	0.894427	1.074011	1.680922
target	4175.0	0.346587	0.475940	0.000000	0.000000	0.000000	1.000000	1.000000

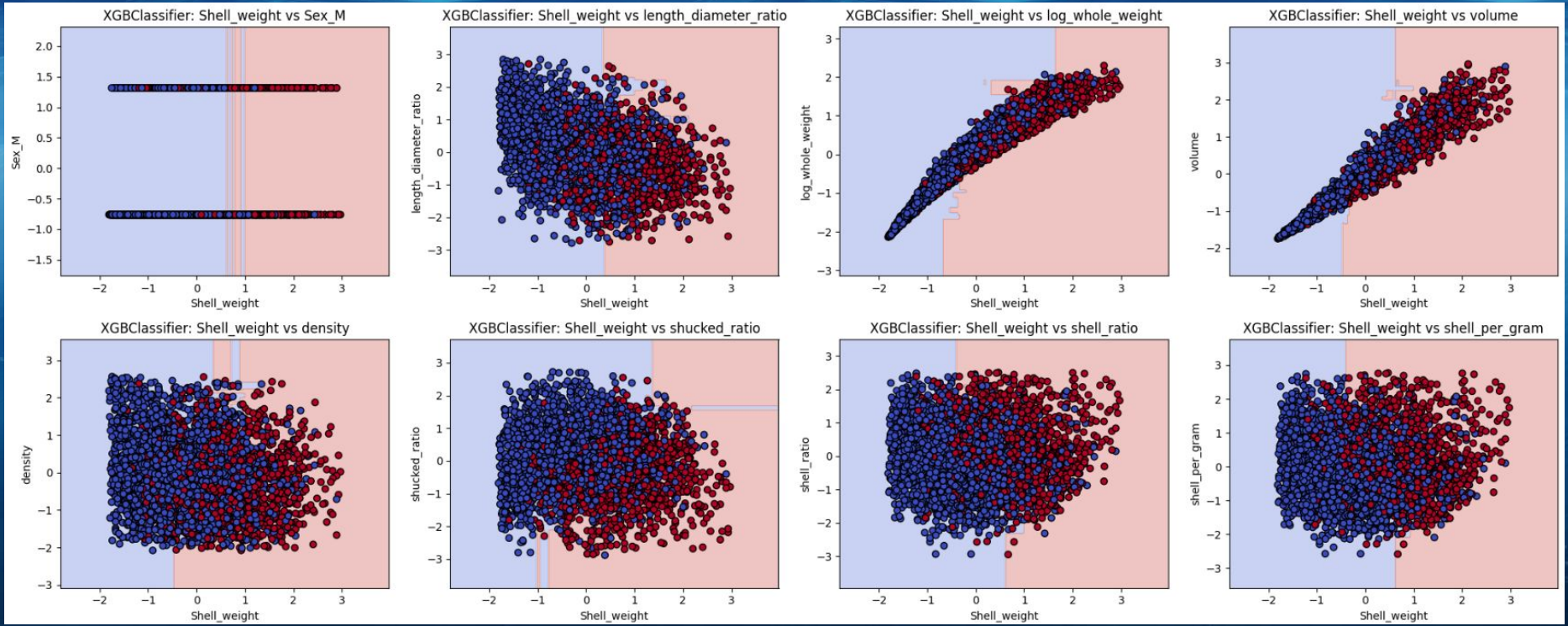
Модели и метрики

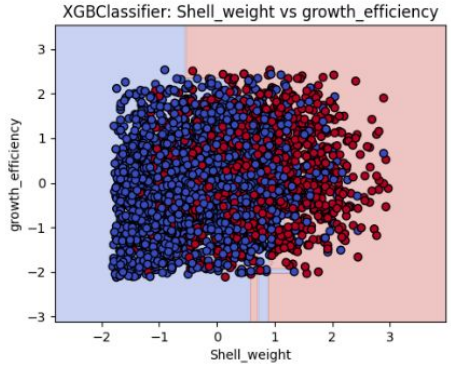
Модель	Accuracy	F1-score	ROC-AUC
XGBoost	0.788 ± 0.011	**0.675 ± 0.020**	**0.863 ± 0.013**
Boosting	0.787 ± 0.007	0.674 ± 0.015	0.860 ± 0.011
CatBoost	0.776 ± 0.013	0.667 ± 0.024	0.850 ± 0.012
Random Forest	0.778 ± 0.015	0.664 ± 0.027	0.852 ± 0.014
Logistic Regression	0.782 ± 0.014	0.647 ± 0.018	0.848 ± 0.012
SVM	0.778 ± 0.014	0.632 ± 0.018	0.830 ± 0.010
Decision Tree	0.707 ± 0.005	0.581 ± 0.004	0.678 ± 0.003
KNN	0.742 ± 0.012	0.579 ± 0.026	0.787 ± 0.009

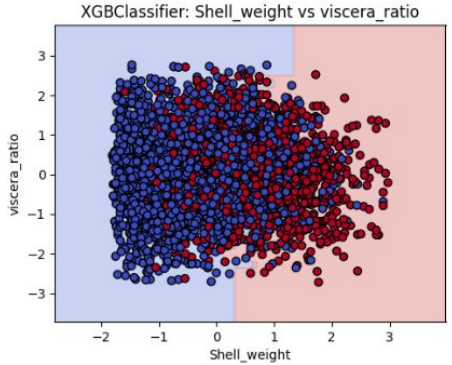
Вывод:
XGBoost — лучшая модель по F1-score и ROC-AUC

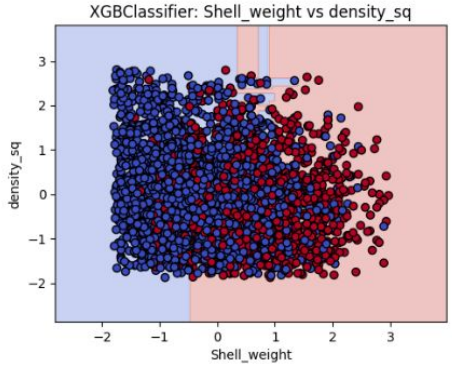
Графическое представление решения модели XGBoost

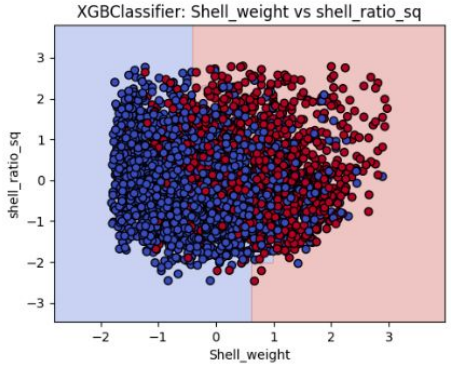


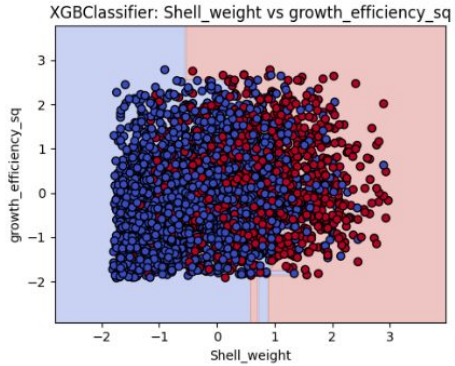

XGBClassifier: Shell_weight vs Sex_M
Sex_M
Shell_weight
XGBClassifier: Shell_weight vs length_diameter_ratio
length_diameter_ratio
Shell_weight
XGBClassifier: Shell_weight vs log_whole_weight
log_whole_weight
Shell_weight
XGBClassifier: Shell_weight vs volume
volume
Shell_weight
XGBClassifier: Shell_weight vs density
density
Shell_weight
XGBClassifier: Shell_weight vs shucked_ratio
shucked_ratio
Shell_weight
XGBClassifier: Shell_weight vs shell_ratio
shell_ratio
Shell_weight
XGBClassifier: Shell_weight vs shell_per_gram
shell_per_gram
Shell_weight

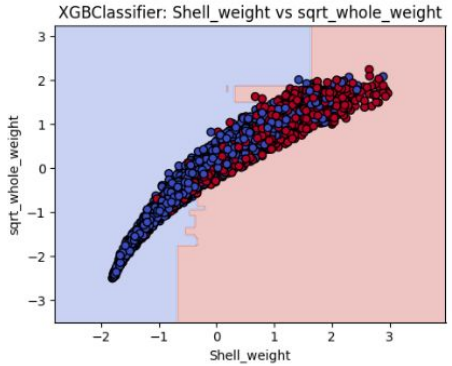

XGBClassifier: Shell_weight vs growth_efficiency
growth_efficiency
Shell_weight


XGBClassifier: Shell_weight vs viscera_ratio
viscera_ratio
Shell_weight


XGBClassifier: Shell_weight vs density_sq
density_sq
Shell_weight


XGBClassifier: Shell_weight vs shell_ratio_sq
shell_ratio_sq
Shell_weight


XGBClassifier: Shell_weight vs growth_efficiency_sq
growth_efficiency_sq
Shell_weight


XGBClassifier: Shell_weight vs sqrt_whole_weight
sqrt_whole_weight
Shell_weight

Калькулятор

- Интерактивный ввод 7 измерений + пол
- Генерация всех 13 признаков
- Предсказание XGBoost с вероятностями

Пример полученных результатов:

Параметр	Значение
Пол (Sex)	M
Длина (Length)	0.5500 см
Диаметр (Diameter)	0.4500 см
Высота (Height)	0.1800 см
Общий вес (Whole)	0.9930 г
Вес мяса (Shucked)	0.3230 г
Вес внутренностей	0.1697 г
Вес раковины (Shell)	0.3000 г

Предсказание: СТАРАЯ (>10 колец)
Вероятность: молодая — 5.6%, старая — **94.4%**

Итог работы

- Ключевые признаки:

  Shell_weight, volume, log_whole_weight — топ-3 по важности

  → сильная корреляция с target и значительный вклад в качество

- XGBoost — лучшая модель:

  F1-score = 0.675, ROC-AUC = 0.863 → устойчиво превосходит другие методы

- Созданы 13 информативных признаков, включая нелинейные преобразования

- Реализован интерактивный калькулятор — практическое применение модели

Модель готова к использованию в биологии и аквакультуре